Enhancing Precision in Rectal Cancer Radiotherapy: Localized Fine-Tuning of Deep-learning based Auto-segmentation (DLAS) Model for Clinical Target Volume and Organs-at-risk

Jianhao Geng
Peking University Cancer Hospital & Institute

Xin Sui
Peking University Cancer Hospital & Institute

Rongxu Du
Peking University Cancer Hospital & Institute

Jialin Feng
Peking University Cancer Hospital & Institute

Ruoxi Wang
Peking University Cancer Hospital & Institute

Meijiao Wang
Peking University Cancer Hospital & Institute

Kaining Yao
Peking University Cancer Hospital & Institute

Qi Chen
MedMind Technology Co., Ltd

Lu Bai
MedMind Technology Co., Ltd

Shaobin Wang
MedMind Technology Co., Ltd

Yongheng Li
Peking University Cancer Hospital & Institute

Hao Wu
Peking University Cancer Hospital & Institute

Xiangmin Hu
Beijing Institute of Technology

Yi Du (yidu_rt@163.com)
Peking University Cancer Hospital & Institute
Enhancing Precision in Rectal Cancer Radiotherapy: Localized Fine-Tuning of Deep-learning based Auto-segmentation (DLAS) Model for Clinical Target Volume and Organs-at-risk

Jianhao Geng, Xin Sui, Rongxu Du, Jialin Feng, Ruoxi Wang, Meijiao Wang, Kaining Yao, Qi Chen, Lu Bai, Shaobin Wang, Yongheng Li, Hao Wu, Xiangmin Hu and Yi Du

1 Key laboratory of Carcinogenesis and Translational Research (Ministry of Education/Beijing), Department of Radiation Oncology, Peking University Cancer Hospital & Institute, Beijing 100142, China.

2 Research and Development Department, MedMind Technology Co., Ltd., Beijing 100083, China.

3 Institute of Medical Technology, Peking University Health Science Center, Beijing 100191, China.

4 Beijing Key Lab of Nanophotonics and Ultrafine Optoelectronic Systems, School of Optics and Photonics, Beijing Institute of Technology, Beijing 100081, China.

5 These three authors contributed equally to this work.
Correspondence: Yi Du, yidu_rt@163.com.

Xiangmin Hu, hu.xiangmin@bit.edu.cn.
Abstract

Background and Purpose

Various deep learning auto-segmentation (DLAS) models have been proposed, some of which commercialized. However, the issue of performance degradation is notable when pretrained models are deployed in the clinic. This study aims to enhance precision of a popular commercial DLAS product in rectal cancer radiotherapy by localized fine-tuning, addressing challenges in practicality and generalizability in real-world clinical settings.

Materials and Methods

A total of 120 Stage II/III mid-low rectal cancer patients were retrospectively enrolled and divided into three datasets: training (n = 60), external validation (ExVal, n = 30), and generalizability evaluation (GenEva, n = 30) datasets respectively. The patients in the training and ExVal dataset were acquired on the same CT simulator, while those in GenEva were on a different CT simulator. The commercial DLAS software was first localized fine-tuned for clinical target volume (CTV) and organs-at-risk (OAR) using the training data, and then validated on ExVal and GenEva respectively. Performance evaluation involved comparing the localized fine-tuned model (LFT) with the vendor-provided pretrained model (VPM) against ground truth contours, using metrics like Dice similarity coefficient (DSC), 95th Hausdorff distance (95HD), sensitivity and specificity.

Results

Localized fine-tuning significantly improved CTV delineation accuracy (p < 0.05)
with LFT outperforming VPM in target volume, DSC, 95HD and specificity. Both models exhibited adequate accuracy for bladder and femoral heads, and LFT demonstrated significant enhancement in segmenting the more complex small intestine. We did not identify performance degradation when LFT and VPM models were applied in the GenEva dataset.

**Conclusions**

The necessity and potential benefits of localized fine-tuning DLAS towards institution-specific model adaption is underscored. The commercial DLAS software exhibits superior accuracy once localized fine-tuned, and is highly robust to imaging equipment changes.

**Keywords**

Deep learning, rectal cancer, CTV, OAR
1. Introduction

Clinical target volume (CTV) and organ-at-risk (OAR) delineation are crucial in successful radiotherapy (RT) treatment planning, with accurate segmentation being vital for delivering safe and effective radiation doses to tumor lesions while minimizing damage to surrounding normal tissues. Conventionally, radiation oncologists manually contour CTV and OAR structures slice by slice, the process of which is substantially labor-intensive and error-prone. Over the past years, considerable efforts have been made towards developing deep-learning based auto-segmentation (DLAS) models specific for CTV and OAR delineation in radiotherapy. Numerous studies demonstrate promising benefits of DLAS in both accuracy and efficacy over atlas-based methods [1-4]. Also, DLAS's effectiveness in mitigating dose inconsistencies has been notably observed in a simulation study [5] based on RTOG 0617 [6, 7], a multi-institutional clinical trial, highlighting substantial potential in streamlining and standardizing clinical workflows. Despite proven advantages in DLAS, most advancements predominantly rely on in-house developed DLAS models, and the widespread implementation of DLAS as a routine tool in clinical settings is far below anticipation. Challenges persist in the clinical adoption of DLAS models, which is highlighted by a survey across 246 institutions where only 26% reported using DLAS in their clinical practice [8].

Generalizability remains a primary challenge for current DL models, where validated models may perform inferiorly in clinical scenarios not represented in their training procedure. For example, deploying top-rated DLAS models from prestigious
challenges directly to local institutions resulted in suboptimal performance [9]. Similarly, several studies [10–12] reported notable performance decline when using DL models in external data. Duan et al [13] evaluated three commercial DLAS products with local cases and observed compromised performance as well.

The mismatching among training and validation datasets, known as data shift, contributes significantly to performance deterioration after clinical deployment of DLAS models [14]. Such shifts may result from variations in clinical practices [15], evolution of delineation guidelines [16], or differences in imaging equipment.

To address these issues, upfront model recalibration or adaption is recommended to meet institution-specific standards previous to clinical application [15]. Instead of retraining DLAS models from the ground, a viable and efficient solution is to incrementally retrain or localized fine-tune pretrained models using pooled data to incorporate institution-specific protocols. Balagopal et al [17] proposed a network model named PSA-Net that segments CTV for postoperative prostate cancer, and observed 5% DSC improvement when adapting to the style of a separate institution.

Fortunately, some commercial vendors offer model retraining services or research tools for users to customize their DLAS models using institution-specific data. However, relevant studies on their clinical implementation is quite limited. Previous studies by Duan et al [13] and Hobbis et al [18] investigated fine-tuning a commercial DLAS software (INTContour, CarinaAI) for OAR structures in prostate cancer patients. Yet, experiences in localized adaptation specifically for CTV or other tumor sites remain unexplored.
To this end, this study addresses this notable gap by detailing the process and outcomes of localized fine-tuning and validation of a popular commercial DLAS product for rectal cancer radiotherapy in a real-world setting. The key novelties and contributions of our work are manifold: (1) first study on DLAS model fine-tuning specifically for rectal cancer radiotherapy, (2) utilization of a popular DLAS product in mainland China with > 400 installations, (3) comprehensive focus on the adaptation and validation of both CTV and OAR structures, and (4) practical insights into model generalizability in the context of changes in imaging equipment—a frequent scenario in clinical settings and we eventually encountered.

2. Materials & Methods
Figure 1. Conceptual design and implementation workflow of this study in model fine-tuning and performance evaluation (external validation and generalizability evaluation).

The conceptual design and overall workflow are shown in Figure 1. The work is generally composed of two procedures-model fine-tuning, and performance evaluation. The latter includes external validation, evaluating model performance on patients scanned on the same CT simulator as training patients but not utilized during
model training. Generalization evaluation refers to assessing the model performance on patients scanned on a different CT simulator and not involved in model training.

2.1 Data Collection

2.1.1 Patient Cohort

This retrospective study was approved by the institutional review board (IRB) of Peking University Cancer Hospital. A total of 120 patients were included in this work, who were diagnosed with Stage II/III mid-low rectal cancer and received chemoradiation at the institutional radiotherapy department. Over the enrolled cohort, 71 were female and 49 were male, and the ages ranged from 33 to 86 with the median as 65.

The enrolled patients were grouped into three datasets - a training dataset, an external validation dataset denoted as ExVal and a generalizability evaluation dataset denoted as GenEva as shown in Table 1. The training dataset was composed of 60 patients treated between March 2020 and October 2022. The external validation dataset ExVal was composed of 30 patients treated between November 2022 and May 2023. At the end of 2022, a Philips RT-specific CT scanner was commissioned into clinical service at our institution, and 30 patients scanned on this CT-Sim between February 2023 and May 2023 were collected as the dataset GenEva to evaluate model generalizability.

Table 1: Description of patient data grouping
<table>
<thead>
<tr>
<th>Dataset Name</th>
<th>Cohort Volume, Age</th>
<th>CT Scanner</th>
<th>Image Acquisition Protocol</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Female/Male (min-max, median)</td>
<td></td>
<td>(kVp, FOV, resolution)</td>
</tr>
<tr>
<td><strong>Training</strong></td>
<td>60, 21/39 33-83, 62</td>
<td>Siemens Sensation Open CT simulator</td>
<td>120, 65 cm, 1.27 x 1.27 x 5 mm³</td>
</tr>
<tr>
<td><strong>External Validation</strong></td>
<td>30, 10/20 37-86, 63</td>
<td>Siemens Sensation Open CT simulator</td>
<td>120, 65 cm, 1.27 x 1.27 x 5 mm³</td>
</tr>
<tr>
<td><strong>Generalization</strong></td>
<td>30, 9/21 41-83, 66</td>
<td>Philips Big-bore RT simulator</td>
<td>120, 60cm, 1.17 x 1.17 x 5 mm³</td>
</tr>
</tbody>
</table>

2.1.2 Image Acquisition

Patients in this study were immobilized by pelvis thermoplastic in a supine posture. The training dataset and ExVal were scanned on a Siemens Sensation Open CT simulator, while the GenEva dataset was scanned on a Philips Big-Bore CT simulator. Detailed specifications of the scan parameters are listed. The CT images were imported into the Eclipse Treatment Planning System (Varian Medical System Inc., USA) for physician to delineate target and OAR structures. The contours as well as plans were reviewed by an internal panel before approved for clinical treatment. In this retrospective study, we retrieved the planning CT images as well as CTV and OAR contours from the treatment planning system in an anonymized approach under the IRB approval. The CTV and OAR contours approved for treatment were used ground truth (GT) reference in model training and performance evaluation. It is
important to note that all the contours were real-world data and we did not do any editing to refine them for the purpose of this study.

2.2 DL Model and Localized Fine-tuning

2.2.1 DL Kernel Network

The DL model for rectal cancer neoadjuvant radiotherapy herein was adopted from the work by Wu et al [19, 20] and commercialized as RT-Mind-AI (MedMind Technology Co. Ltd., Beijing, China). The backbone network, referred as DpnUNet, was characterized by integrating dual-path-network (DPN) modules into the UNet structure. The overall architecture of DpnUNet was generally depicted in Figure 2.

![Figure 2. Schematic of the kernel DpnUNet network architecture](image)

2.2.2 Localized Model Fine-tuning

The model was pretrained using 122 patients’ data from a single institution [19]. We further trained the model with the enrolled training data (60 patients) to adapt to the
institutional contouring protocol. The contours of interest were CTV, bladder, femoral heads and small intestine. The class weighted cross-entropy was used to take into account the overall accuracy in both CTV and OARs. Localized model fine-tuning was performed on a single GPU workstation (Nvidia GeForce RTX 2080Ti) using 5-fold cross validation (48 vs. 12). The optimizer was Adam, and the batch size was 4. The initial learning rate was 0.0001, and the value decayed exponentially by a factor of 0.9 over each epoch. The epoch was 60, and the model with the lowest cross-validation loss was selected as the final output.

2.3 Performance Evaluation

2.3.1 External Validation and Generalizability Evaluation

This study used two datasets (ExVal and GenEva) with 30 cases in each to evaluate model performance in two aspects. The data in ExVal were acquired on the same CT simulator with the training data, and therefore used for external validation. The data in GenEva were acquired on a different CT simulator, and herein were used to evaluate model generalization in the context of imaging equipment changes.

2.3.2 Quantitative metrics

Two sets of deep learning predicated contours were generated for all 60 testing cases, using a vendor-provided pretrained model (VPM) and a localized fine-tuned model (LFT) respectively. We utilized several valid and widely used metrics to quantify segmentation performance, including the Dice Similarity Coefficient (DSC), the 95th percentile of the Hausdorff Distance (95HD), sensitivity, and specificity, using the
clinically approved CTV and OAR contours as GT.

DSC, the most used measure in the field of medical image segmentation, provides an effective assessment of similarity, and is defined as:

\[
DSC(D, G) = \frac{2|D \cap G|}{|D| + |G|} \quad (6),
\]

where \( D \) and \( G \) represent the DLAS-predicted and GT contours respectively, and \(|D \cap G|\) represents the intersected volume between \( D \) and \( G \).

The 95HD metric is a routinely used spatial distance-based metric to measures the distance between the DLAS-predicted and GT contours, which is defined as

\[
95HD(D, G) = \text{percentile}(h(D, G) \cup h(G, D), 95th) \quad (7),
\]

\[
h(D, G) = \max_{d_i}(\min_{g_j}||d_i - g_j||), d_i \in D, g_j \in G \quad (8),
\]

where \(||.||\) stands for the Euclidean norm of the points of \( d \) and \( g \).

Sensitivity and specificity are popular metrics for the evaluation of medical image segmentation performance [14-15], which are defined as

\[
\text{Sensitivity} = \frac{TP}{TP + FN} \quad (9),
\]

\[
\text{Specificity} = \frac{TN}{TN + FP} \quad (10),
\]

which TP, FP, TN and FN denote the pixel numbers of true positive, false positive, true negative and false negative respectively for DLAS-predicted contours.

In addition, the CTV volume was also measured to meet the research interests of our oncologist team. The DSC, 95HD, sensitivity, specificity, and CTV-volume values of each testing case were calculated in the 3D Slicer software (version 5.4.0) [16].

2.3.3 Statistical Analysis

The mean and standard deviation (SD) values were calculated for each metric. Within
each testing dataset, the Wilcoxon paired signed-rank test was used to compare the performance between VPM and LFT. The statistical analysis was performed in OriginPro (version 2021a, OriginLab, USA), and the significance level was set at 0.05.

3. Results
3.1 Visualization of representative cases

Figure 3 shows representative patient cases that are selected from ExVal and GenEva datasets for subjective visual illustration. Images in ExVal (Fig 3(a-f)) exhibit finer texture patterns, while images in GenEva (Fig 3(g-l)) appear smoother with reduced noise. Also, we can see that CTV and OAR contours predicted either by VPM or LFT models are generally consistent with GT, especially in organs of bladder and femoral heads, where the contours of VPM, LFT and GT are highly overlapped. As for small intestine, while major deviations exhibit in areas where intestine tubes are separate but close to each other as shown in Fig.3(a)&(d)&(j), most intestine loops predicated by VPM and LFT are highly conformal to GT.

3.2 Quantitative assessment of CTV

From Figure 3, we can also see that, although both VPM and LFT contours are generally consistent with GT, CTV volumes predicated by VPM appear to be larger...
than either GT or LFT. This *over-contouring* effect is observed in both *ExVal* and *GenEva* datasets and further validated in case-by-case comparison in Figure 4.

The CTV volumes (mean±std) predicted by VPM and LFT in *ExVal* are 720.304±90.789 cm$^3$ and 606.443±90.677 cm$^3$ respectively with the benchmark GT as 617.879±110.506 cm$^3$. The corresponding relative errors in comparison with GT are 17.921±11.663 % and -1.281±5.655 %. The volume difference between VPM and LFT predicted CTV contours is statistically significant (p-value < 0.05) in *ExVal*. The CTV volumes predicted by VPM and LFT in *GenEva* are 735.997±109.678 cm$^3$ and 610.804±74.917 cm$^3$ respectively with the benchmark GT as 630.320±82.261 cm$^3$. The corresponding relative errors in comparison with GT are 16.923±9.661 % and -2.798±6.228 %. The volume difference between VPM and LFT predicted CTV contours is statistically significant (p-value < 0.05) in *GenEva*. Figure 5 shows the distributions of DSC, sensitivity, specificity and 95HD values for DLAS-predicted CTV contours. It is obvious that LFT model prevails over VPM in metrics of DSC, specificity and 95HD in both *ExVal* and *GenEva* datasets. Specifically, the improvements of DSC mean values are 11.406 % in *ExVal* and 9.340 % in *GenEva* with statistical significance (p-value <0.01). The mean specificity values are improved by 2.497 % in *ExVal* and 2.591 % in *GenEva* (p-value < 0.01), and the mean 95HD values are reduced by 46.866% and 42.120% respectively (p-value < 0.01). On the contrary, the sensitivity distributions between VPM and LFT predicted CTV contours are very close in both *ExVal* and *GenEva* (p-value > 0.05).
Figure 4. Profile and relative errors of VPM and LFT predicted CTV volumes in comparison with GT over (a) ExVal, and (b) GenEva respectively. (GT-black line, VPM-red line/bar, and LFT-blue line/bar)
Figure 5. Distribution comparison and statistical analysis of the CTV contours by VPM and LFT models in (a) ExVal and (b) GenEva compared with GT in metrics of DSC, sensitivity, specificity and 95HD respectively.

### Table 2

<table>
<thead>
<tr>
<th>Metrics</th>
<th>VPM</th>
<th>LFT</th>
</tr>
</thead>
<tbody>
<tr>
<td>DSC</td>
<td>0.789</td>
<td>0.789</td>
</tr>
<tr>
<td>Sensitivity</td>
<td>0.879</td>
<td>0.879</td>
</tr>
<tr>
<td>Specificity</td>
<td>0.877</td>
<td>0.877</td>
</tr>
<tr>
<td>95HD (mm)</td>
<td>19.099</td>
<td>19.099</td>
</tr>
</tbody>
</table>

* p < 0.01

3.3 Quantitative assessment of OAR

Table 2 summarizes the statistical analysis in metrics of DSC, 95HD, sensitivity...
and specificity for OAR structures. As for bladder and femoral heads, both VPM and LFT models exhibit sufficient and comparable accuracy. Despite that some p-values are <0.05, differences in metrics values between LFT and VPM, \textit{ExVal} and \textit{GenEva} are small as negligible. This is consistent with the overlapping contours in Figure 3, which demonstrates that both VPM and LFT models are adequately accurate in segmenting bladder and femoral heads. As for small intestine, we can see that LFT generally outperformed VPM (p-value <0.05 in DSC-\textit{GenEva}, 95HD-\textit{GenEva}, sensitivity-\textit{ExVal} and sensitivity-\textit{GenEva}) except in specificity, of which the difference is negligible.

\textbf{Table 2.} Summary and statistical analysis of OAR structures (bladder, femoral heads, and small intestine) predicated by VPM and LFT in comparison with GT in metrics of DSC, 95HD, sensitivity and specificity.

<table>
<thead>
<tr>
<th></th>
<th>DSC</th>
<th>95HD (mm)</th>
<th>Sensitivity</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>\textit{ExVal}</td>
<td>\textit{GenEva}</td>
<td>\textit{ExVal}</td>
<td>\textit{GenEva}</td>
</tr>
<tr>
<td>Bladder</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>VPM</td>
<td>0.967±0.013</td>
<td>0.964±0.015</td>
<td>2.837±1.465</td>
<td>3.311±1.723</td>
</tr>
<tr>
<td>LFT</td>
<td>0.961±0.011</td>
<td>0.961±0.014</td>
<td>3.327±1.065</td>
<td>3.582±1.881</td>
</tr>
<tr>
<td>p-value</td>
<td>&lt;0.01</td>
<td>&lt;0.01</td>
<td>0.32</td>
<td>0.51</td>
</tr>
<tr>
<td>Femoral Heads</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>VPM</td>
<td>0.956±0.013</td>
<td>0.972±0.011</td>
<td>2.193±0.573</td>
<td>1.922±0.736</td>
</tr>
<tr>
<td>LFT</td>
<td>0.949±0.008</td>
<td>0.970±0.013</td>
<td>2.478±0.532</td>
<td>1.978±0.515</td>
</tr>
<tr>
<td>p-value</td>
<td>0.01</td>
<td>0.14</td>
<td>0.03</td>
<td>0.62</td>
</tr>
<tr>
<td>Small Intestine</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>VPM</td>
<td>0.838±0.072</td>
<td>0.806±0.068</td>
<td>8.312±2.821</td>
<td>11.605±3.153</td>
</tr>
<tr>
<td>LFT</td>
<td>0.853±0.051</td>
<td>0.834±0.084</td>
<td>8.201±3.229</td>
<td>10.888±3.430</td>
</tr>
<tr>
<td>p-value</td>
<td>0.13</td>
<td>&lt;0.01</td>
<td>0.91</td>
<td>0.03</td>
</tr>
</tbody>
</table>
4. Discussion

In this study, we meticulously detailed our process and outcomes from localized fine-tuning and validation of a popular commercial DLAS product, RT-Mind-AI, specifically targeting rectal cancer radiotherapy. This work marks a significant stride in not only addressing the imperative need in enhancing DLAS model performance in real-word clinical settings but also the generalizability of RT-Mind-AI in the context of imaging equipment changes.

In the process of localized model fine-tuning, we used real-world patient data that had been approved for clinical treatment with minimal data preprocessing. The training cohort was composed of 60 patients, much smaller than pertinent studies on in-house DLAS model development (135 patients in [19], 218 patients in [21], 136 patients in [22], 100 patients in [23]) and on incrementally training a commercial DLAS model in [13] (with 100 patients) as well. The use of a small volume of real-world patient data has underscored the practicality and efficiency of on-site data preparation, which facilitates the localized fine-tuning process without extensive data collection and labor-intensive data preprocessing.

The CTV evaluation results demonstrate that the fine-tuned RT-Mind-AI exhibited comparable performance with in-house models (mean DSC = 0.879 in Exval/0.874 in GenEva, 0.90 in [19], 0.877 in [21], 0.88 in [22], 0.90 in [23]). Notably, the LFT model showed superior performance over the VPM model, especially in the accuracy of CTV volume estimation and the reduction of over-contouring tendencies. The
substantial improvements in metrics such as DSC and 95HD highlight the effectiveness of adapting models to institution-specific clinical standards.

For bladder and femoral heads, the segmentation performance of both VPM and LFT was adequately accurate, as indicated by the high metrics values. This finding obviously indicates that the vendor-provided model has been well trained and these model components may not require further retraining, potentially easing the implementation process in clinical settings. This indication is consistent with the study by [18]. This is mainly due to the distinct anatomical characteristics of bladder and femoral heads and lower inter-observer variability among physicians. Accurate segmentation of the small intestine poses a significantly greater challenge than that of the bladder and femoral heads. Although the LFT method has shown significant improvements in DSC, 95HD and sensitivity metrics (p-value < 0.05), these metrics still fall short of the benchmarks established in bladder and femoral heads segmentation. This discrepancy is primarily due to two factors. Firstly, the small intestine's anatomy is inherently complex, exhibiting considerable variability in positioning and filling of the intestines, which makes distinguishing between intestinal loops and adjacent tissues particularly difficult. Secondly, the variation in contouring the small intestine among different observers is significant [13, 18], making it challenging to achieve a strictly consistent standard. This also highlights the importance of enhancing DLAS accuracy for complex anatomical structures.

Equipment changes are significant events that may take place occasionally in clinical practice, and one of the significant aspects of this study is addressing the
impact of different CT simulators on model performance. The robustness of the commercial DLAS product, even with changes in imaging equipment, is a promising finding for institutions undergoing technological upgrades or those using multiple imaging systems. This robustness is crucial for the widespread adoption of DLAS technologies, ensuring consistent performance across various clinical environments.

Despite these promising results, we acknowledge several limitations in this work. First, this study is based on a specific commercial DLAS product (RT-Mind-AI) and two CT simulators. As for other DLAS products or simulators, some findings may not be solidly valid. The generalizability of these results to other DLAS products or imaging equipment needs further investigation. Second, the cohort size for model fine-tuning was empirical based on our previous work [24]. Maybe a smaller cohort would be enough, and future studies might explore the optimal cohort size for effective and efficient model fine-tuning. Third, this study is focused on rectal cancer radiotherapy. Next-step efforts will be directed to DLAS model fine-tuning for other tumor sites to facilitate clinical application as well as further validation.

5. Conclusion

We detailed the process and outcomes from localized fine-tuning and validation of a popular commercial DLAS product (RT-Mind-AI) specifically for rectal cancer radiotherapy in real-world clinical settings. The comprehensive validation explicitly underscores the necessity and potential benefits in institution-specific DLAS model adaption and continues model updating, which indicates that localized model
fine-tuning for various clinical settings is crucial in realizing the full potential of
DLAS in enhancing the precision and effectiveness of radiotherapy treatments. This
work also demonstrate that the RT-Mind-AI software is highly robust to imaging
equipment changes, and exhibits superior accuracy once localized fine-tuned.

Abbreviations

95HD: 95th percentile Hausdorff distance
CNN: convolutional neural network
CT: computed tomography
CTV: clinical target volume
DL: deep-learning
DLAS: deep learning auto-segmentation
Dpn: dual-path network
DSC: Dice similarity coefficient
GT: ground truth
GTV: gross tumor volume
LFT: localized fine-tuned
OAR: organ-at-risk
ROI: region of interest
RT: radiotherapy
SD: standard deviation
VPM: vendor-provided pretrained
Declarations

Ethics approval and consent to participate

The study was approved by the IRB of Peking University Cancer Hospital, China.

Consent for publication

All the authors reviewed the manuscript and approved the final submission to Clinical and Translational Radiation Oncology.

Availability of supporting data

The data that support this study are not openly available due to ethical and privacy concerns but are available from the corresponding author upon reasonable request.

Competing interests

The authors have no conflicts of interest.

Funding

This work was supported in part by National Natural Science Foundation of China (No. 12375335, 12005007), National Key Research and Development Project (No. 2019YFF01014405), Youth Fund of the National Natural Science Foundation of China (No.22305015), Clinical Research Project of Wu Jieping Medical Foundation (No. 320.6750.2022-3-59) and Peking University Cancer Hospital Science Foundation JC20240.

Authors’ contributions

JG, XS and RD contributed equally to this paper. Conceptualization & Data curation & Writing (original draft): JG, RD and JF. Methodology & Formal Analysis &
References

https://doi.org/10.1186/s13014-020-01617-0.


https://doi.org/10.1002/mp.14320.


